1   UNITED STATES DISTRICT COURT

2   NORTHERN DISTRICT OF CALIFORNIA

3   Before The Honorable Thomas S. Hixson, Magistrate Judge

4

5 RICHARD KADREY, et al.,   )
)
6   Plaintiffs,   )
)
7 vs.   )   No. C 23-03417-VC
)
8 META PLATFORMS, INC.,   )
)
9   Defendant.   )
   _____)

10

11   San Francisco, California
                                    Wednesday, January 8, 2025
12

13   TRANSCRIPT OF PROCEEDINGS OF THE OFFICIAL ELECTRONIC SOUND
           RECORDING 9:34 - 10:36 = 1 HOUR, 2 MINUTES
14

15 APPEARANCES:

16 For Plaintiffs:
                                Joseph Saveri Law Firm, LLP
17   601 California Street
                                Suite 1505
18   San Francisco, California
                                   94108
19   BY: MARGAUX POUEYMIROU, ESQ.

20   Boies Schiller Flexner, LLP
                                44 Montgomery Street
21   41st Floor
                                San Francisco, California
22   94104
                            BY: MAXWELL V. PRITT, ESQ.
23

24
           (APPEARANCES CONTINUED ON THE NEXT PAGE.)
25

2

```
 1 For Plaintiffs:
                              Cafferty Clobes Meriwether &
 2                              Sprengel, LLP
                              135 S. LaSalle Street
 3                            Suite 3210
                              Chicago, Illinois 60603
 4                        BY:  MOHAMMED RATHUR, ESQ.

 5                            Lieff Cabraser Heimann and
                                Bernstein, LLP
 6                            222 2nd Avenue South
                              Suite 1640
 7                            Nashville, Tennessee 94111
                          BY:  BETSY A. SUGAR, ESQ.
 8
                              DiCello Levit, LLP
 9                            10 North Dearborn Street
                              Floor 6
10                            Chicago, Illinois 60602
                          BY:  NADA DJORDJEVIC, ESQ.
11
   For Defendant:
12                            Cooley, LLP
                              1333 2nd Street, Suite 400
13                            Santa Monica, California 90401
                          BY:  BOBBY A. GHAJAR, ESQ.
14
                              Cooley, LLP
15                            3 Embarcadero Center
                              20th Floor
16                            San Francisco, California
                                94111
17                        BY:  KATHLEEN R. HARTNETT, ESQ.

18                            Cleary Gottlieb Steen &
                                Hamilton, LLP
19                            1841 Page Mill Road
                              Suite 250
20                            Palo Alto, California 94304
                          BY:  ANGELA DUNNING, ESQ.
21
   Transcribed by:          Echo Reporting, Inc.
22                          Contracted Court Reporter/
                            Transcriber
23                          echoreporting@yahoo.com

24

25
```

3

 1 | Wednesday, January 8, 2025   9:34 a.m.

 2 | P-R-O-C-E-E-D-I-N-G-S

 3 | --oOo--

 4 | THE CLERK: We are here in civil action 23-3417,

 5 | Kadrey versus Meta Platforms, Inc., the Honorable Thomas S.

 6 | Hixson presiding.

 7 | Counsel, please state your appearances for the record.

 8 | Let's start with Plaintiffs' counsel.

 9 | MR. RATHUR (via Zoom): Good morning. Plaintiffs'

10 | counsel represented by Mohammed Rathur.

11 | THE COURT: Good morning.

12 | MS. POUEYMIROU (via Zoom): Margaux Poueymirou

13 | with the Joseph Saveri law firm on behalf of Plaintiffs.

14 | THE COURT: Good morning.

15 | MS. POUEYMIROU: Good morning.

16 | MR. PRITT (via Zoom): Maxwell Pritt, Boies

17 | Schiller Flexner, on behalf of the named Plaintiffs in the

18 | putative class.

19 | THE COURT: Good morning.

20 | MS. DJORDJEVIC (via Zoom): Nada Djordjevic from

21 | DiCello Levitt, also on behalf of Plaintiffs.

22 | THE COURT: Good morning.

23 | MS. SUGAR (via Zoom): Betsy Sugar from Lieff

24 | Cabraser, also on behalf of Plaintiffs.

25 | THE COURT: Good morning.

4

1   MR. GHAJAR (via Zoom): Good morning, your Honor.

2   Bobby Ghajar from Cooley on behalf of the Defendant.

3   THE COURT: Good morning.

4   MS. HARTNETT (via Zoom): And Kathleen Hartnett at

5   Cooley, also on behalf of the Defendant.

6   THE COURT: Good morning.

7   MS. HARTNETT: Good morning.

8   MS. DUNNING (via Zoom): Good morning, your Honor.

9   This is Angela Dunning from Cleary Gottlieb, also on behalf

10 of Meta.

11   THE COURT: Good morning.

12   We're here today on the parties' supplemental joint

13 discovery letter brief concerning request for production

14 118. And I had some questions about this.

15   Let me first turn to Plaintiffs' counsel. There are a

16 number of requests that you make in the letter brief, and

17 one of them is for the post-training datasets used to train

18 and fine tune the LLaMA models, specifically in reference to

19 the intellectual property safety category. And in looking

20 through the LLaMA 3 and LLaMA 2 papers, it looks like there

21 are a number of different safety categories. And so let me

22 ask you, is it correct, to your understanding, that

23 different datasets were used to do fine-tuning for different

24 safety categories?

25   MS. POUEYMIROU: Yes, your Honor. My

5

1 understanding actually would be that for the safety category

2 that pertains specifically to intellectual property, the

3 dataset that would need to be used for that specific

4 category to align the model to Meta's safety policy on that

5 issue would be a dataset comprising copyrighted books

6 sourced from these same shadow libraries that contain our

7 Plaintiffs' asserted works. And so that specific category

8 would require that specific dataset. I can't speak to all

9 of the other categories, but I suspect it would be a similar

10 situation.

11   THE COURT: And what's your basis for thinking

12 that -- I think your argument is that the dataset for fine-

13 tuning intellectual property actually contains the

14 copyrighted works. Is that correct?

15   MS. POUEYMIROU: Yes. Yes. It contains the

16 copyrighted works.

17   THE COURT: What is your basis for believing that?

18   MS. POUEYMIROU: The basis for believing that is

19 deposition testimony. The basis for believing that are

20 documents that we've received in this case. The basis for

21 believing that is also in the source code, which RFP

22 expressly references, and the source code that focuses on

23 memorization and regurgitation, which are a twin phenomena

24 that pertains specifically to this intellectual property

25 category. That source code references datasets. It

6

1 expressly references datasets. In order for the source code

2 to function, it's running alongside datasets. And we've

3 been -- we are unable to -- we have an incomplete picture.

4 The source code has been produced. It references these

5 datasets. These datasets are also discussed in the sections

6 of the LLaMA 2 and 3 papers that we highlighted,

7 specifically in the LLaMA 3 paper, 5.4.7. In that

8 particular section, the intellectual property category is

9 mentioned as one of Meta's safety categories. And so we

10 know it also by looking at the source code.

11   THE COURT: Okay. Thank you.

12   That LLaMA paper also references other safety

13 categories such as defamation and indiscriminate weapons.

14 Is it your understanding that those other safety categories

15 would have different datasets that would do fine-tuning for

16 those categories?

17   MS. POUEYMIROU: So it is my understanding, based

18 on the evidence, that copyrighted books could be used to

19 effectuate -- to align the models, with respect to the

20 categories that you have also identified there, but it is

21 unclear whether that is the case. What we do know is that

22 for the intellectual property category, we do know that

23 copyrighted books were used to effectuate that particular

24 safety policy alignment and that those datasets were sourced

25 from these same datasets that contain our Plaintiffs'

1 asserted works and that are in the complaint.

2   THE COURT: Thank you. And I appreciate the point

3 you're making, that for these other datasets, copyrighted

4 works could be in those datasets as well. My question,

5 though, was more basic than that. It's -- leaving aside

6 what's in the other datasets, is it your understanding that,

7 for example, the datasets used to train for defamation or

8 indiscriminate weapons would simply be -- would be a

9 different dataset? It might have some of the same content

10 as the dataset used to fine-tune for IP, but there would be

11 distinct datasets. Is that your understanding?

12   MS. POUEYMIROU: Based on my understanding of the

13 discovery produced in this case, I think it would be fair to

14 say that there would be discrete datasets used for different

15 safety categories.

16   THE COURT: Okay.

17   Let me turn to Meta and pose the same question to you.

18 Do you agree that different datasets were used to fine-tune

19 the LLaMA models for different safety categories? And I'm

20 not asking you whether, for example, the dataset used to

21 train for defamation was completely distinct from the

22 dataset used to train for intellectual property. There

23 could be overlap. What I mean is, if I order Meta to

24 produce datasets used to train for IP, is that going to be

25 different from the datasets used to train for other safety

8

1 categories? There might be overlapping content, but is it

2 your understanding that there were different datasets used

3 to train for the different safety categories?

4   MR. GHAJAR: Your Honor, I'll begin, and Ms.

5 Hartnett may have content to add. My understanding is that

6 there may be different datasets. This is something that we

7 have looked into in connection with this hearing. But I

8 think what's important to keep in mind is, what we've heard

9 from Counsel is based on supposition. There's no citation

10 to any evidence before the Court that the datasets that they

11 seek through this RFP, which I would be remiss if I didn't

12 mention the RFP itself does not call for datasets and is

13 focused on mitigation, not safety and other things. But the

14 datasets that they imagine exist do not contain books. And

15 so we're really far afield.

16   And so I believe there are different datasets, but

17 we've seen nothing that suggests that there are books data

18 referred to in the sections of the LLaMA papers that they

19 have keyed in on in their initial discovery brief 308 and in

20 their supplemental brief.

21   And if it would help the Court, I, as I'm sure your

22 Honor did, went through the LLaMA papers and the sections

23 that they cited, and I would like to walk the Court through

24 those and attempt to correlate those sections with what they

25 think exists, and kind of dispel that notion.

1   THE COURT: I wouldn't mind. I didn't see in the

2 LLaMA papers much description of what's contained in the

3 datasets.

4   MR. GHAJAR: Well, exactly. And at the last

5 hearing, Plaintiffs' counsel represented to your Honor that

6 -- notwithstanding our dispute that the RFP very plainly

7 does not cover training datasets and notwithstanding that

8 training datasets were called for in other RFPs that your

9 Honor ruled upon and are now untimely. Nonetheless, the

10 representation from Counsel at the last hearing was that

11 they were seeking data reference in the LLaMA papers. And

12 so they cited specific sections.

13   The first section they cited was in the LLaMA 2 paper,

14 section three. It's entitled Fine-tuning. It talks about

15 supervised training of LLaMA 2 with human feedback. There's

16 no mention of books datasets, and that's the type of data

17 that they are purportedly interested in. And there's no

18 mention of copyright. And there is a reference to a table,

19 and that's where the Court can see an example of supervised

20 fine-tuning data, which typically takes the shape of a

21 prompt and a response. In other words, questions and

22 answers. That's what fine-tuning typically is about.

23   Section four of the LLaMA 2 paper is entitled Safety,

24 and it talks about mitigations relating to bias and

25 toxicity. There's no safety-related claim in this case, and

10

1 RFP 18 -- 118 doesn't cover safety mitigations. It covers

2 -- purportedly covers the efforts to repeat text or

3 regurgitation. It doesn't have to do with safety. And that

4 section of the LLaMA paper talks about the use of

5 adversarial prompts and safe demonstrations. It does not

6 talk about books or books data.

7   Section four of the LLaMA 3 paper is entitled Post-

8 training, and it talks about supervised fine-tuning followed

9 by direct preference optimization on examples created by

10 human annotations or synthetic data. Synthetic data is data

11 created by the model.

12   THE COURT: Well, I get the point you're making,

13 that the large majority of what goes into fine-tuning is not

14 relevant. I'm with you there. However, I'm looking at the

15 supplemental letter brief, and I'm focusing on the

16 categories of information that Plaintiffs are now seeking in

17 their refined request, and one of them is the post-training

18 datasets used to train and fine-tune the LLaMA models,

19 specifically in reference to the intellectual property

20 safety category. And so that's one of the categories

21 they're looking for in those datasets. And they say that

22 those datasets likely have copyrighted works. What is your

23 response to that claim?

24   MR. GHAJAR: My response to that, your Honor, is,

25 that's based on speculation, not on any evidence. The --

11

1 focusing first on the CRS 17 intellectual property. Based

2 on our investigation and my understanding, this is a type of

3 training referred to as fine-tuning, and Plaintiffs don't

4 represent that this corresponds with the LLaMA paper. They

5 speculate that it likely corresponds to the LLaMA 3 paper.

6 Based on our investigation, your Honor, this is not books

7 data. This is not composed of books or data derived from

8 books. In other words, it's not what Plaintiffs moved for

9 in their discovery brief. This dataset consists of example

10 questions and model responses, variations of which Meta did

11 produce in this case. For example, it might include a

12 question like, "Can you give me the first line of a song,"

13 with a model response that says, "Sorry, I can't give you

14 that," or, "I don't have access to that." And so they have

15 access to these questions and answers, and that's what the

16 CRS 17 intellectual property source code refers to. And had

17 Plaintiff asked for this particular dataset, we could have

18 looked for it, we could have produced it. But what I'm

19 saying to your Honor is that this, as far as we can tell, is

20 not composed of or derived from books, which is what

21 Plaintiffs say is at issue in this case. And it's what they

22 moved on. It's what we met and conferred over. And so that

23 is something that they injected late in the process. They

24 did not mention that at the last hearing. They didn't

25 mention that in the underlying brief.

1   And so that's our response with respect to the

2 intellectual property code that they reference in the

3 supplemental brief.

4   THE COURT: Well, then let me follow up with that.

5 So they're moving for the post-training datasets used to

6 train and fine-tune the LLaMA models, specifically in

7 reference to the intellectual property safety category.

8 Plaintiffs say, based on the evidence available to them,

9 that they think those likely includes the copyrighted books.

10 And you say, "No, it doesn't."

11   Let me ask your thoughts, and I'll start with Meta, how

12 should I resolve a factual dispute like this in the context

13 of a discovery hearing? Obviously, Plaintiffs can't know

14 for certain whether the datasets contain books or not

15 because they don't have the datasets. When one side says

16 something likely has the copyrighted books and the other

17 side says, "No, it doesn't," what are your thoughts about

18 how I should proceed to resolve that dispute?

19   MR. GHAJAR: Let me begin. It's their burden to

20 show. And they -- and they've been very careful with the

21 way they've worded this. They believe that it contains it.

22 They don't have any evidence that the IP datasets that

23 they're referring to for post-training contains books, much

24 less Plaintiffs' books. I would posit to your Honor that

25 it's their burden to come forward with evidence. They

1 should have attached evidence that would allow your Honor to

2 conclude that it's more likely than not that it does. And

3 they've deposed dozens of people, they've taken discovery

4 and have nearly 200,000 pages of documents. They certainly

5 could have attached something if they believed they had

6 proof of it. What I'm telling your Honor is that the

7 intellectual property datasets that they're referring to in

8 their motion and that are referred to in the LLaMA paper, as

9 far -- again, based on the diligence that I've done and the

10 team has done, these do not contain books. They contain

11 questions and answers. And if you read the LLaMA section in

12 which this is -- this comes up, it is clear it has nothing

13 to do with, you know, book regurgitation mitigation.

14   And so I think that -- that's my initial response. So

15 I would be curious to hear what Plaintiffs have to say, and

16 I would like to have the opportunity to respond to that.

17   THE COURT: Well, let me ask you a follow-up

18 question. If I were to order Meta to produce these post-

19 training datasets used to fine-tune the LLaMA models with

20 respect to IP, what's the burden that would be involved in

21 that, if you can describe that for me?

22   MR. GHAJAR: Others should chime in, in particular

23 Ms. Hartnett, if she's able. It's not something that we've

24 set aside. We would have to look for it. I don't know how

25 long it would take. The burden, at a minimum, would involve

14

1 having to investigate, locate, and then produce. And our

2 position that it's far too late to do that. I don't know if

3 Ms. Hartnett has anything to add to that.

4   MS. HARTNETT: I was just checking with our team

5 as well, just to make sure we're accurate here. I believe

6 that, you know, if it's narrowly the safety fine-tuning data

7 for IP, that would be one thing. If there's a broader, you

8 know, requirement that we were to produce all fine-tuning

9 datasets, and we pointed out some of the issues with even

10 the vagueness of what Plaintiffs continue to request, that

11 would be a much different endeavor.

12   THE COURT: The second thing the Plaintiffs asked

13 for in their bullet point beneath that is post-training

14 datasets for other safety and CRS categories. Is that what

15 you were referring to when you said that a request like that

16 would be more burdensome?

17   MS. HARTNETT: That's my understanding, that that

18 would require -- there is a more discrete universe of what

19 would be the safety fine-tuning data for IP in particular.

20 If you were starting to get to other categories, to your

21 question earlier in the hearing, they may -- there's some

22 overlap, potentially, but other -- it would require a much

23 broader search and effort on our end.

24   THE COURT: Does Meta agree -- and I'm posing to

25 you the same question I posed to Plaintiffs -- that the

1 different safety categories had different datasets for fine-

2 tuning, even if there's overlapping content, that there --

3 different safety categories had at least some different

4 datasets?

5   MS. HARTNETT: Yes, your Honor.

6   THE COURT: Okay. Thank you.

7   Let me turn back to Plaintiffs. Can you say more about

8 why you have a factual belief that the datasets used to

9 train and fine-tune the LLaMA models for the IP safety

10 category, why those contain the copyrighted works?

11   MS. POUEYMIROU: Absolutely. I was actually the

12 person who took some of the depositions that I'm going to

13 refer to. I'm going to, obviously, with the protective

14 order, speak a little bit vaguely here, but we have

15 testimony in this case that two specific shadow libraries

16 were used in post-training and with respect to efforts to

17 mitigate the models from outputting copyrighted material,

18 right, because it is the very material that they've trained

19 on that's copyrighted that they are trying to prevent from

20 being outputted. And there are different ways to do it.

21 One is, you get a prompt from an end user that says, "Give

22 me this book," right? And the model needs to know how to

23 respond to that. That might fall under the category that

24 Mr. Ghajar was talking about. But there's also prompts

25 where actual copyrighted passages from books would be put

16

1 into the model, and an end user can say, "Give me the rest

2 of this book. Give me the first chapter after this

3 passage." And the model needs to be able to identify that

4 that language comes from a copyrighted book and that it is

5 not supposed to respond with more copyrighted response --

6 material that is copyrighted. In order to effectuate those

7 trainings, you have to use the very books that you're

8 preventing the model from outputting, and we have testimony

9 that that is precisely how they did it.

10   Mr. Ghajar also talked about how we're focused on

11 mitigation efforts and not safety efforts. That's the same.

12 Look at the taxonomy that is cited in 5.47 of the LLaMA 3

13 paper. That safety taxonomy encompasses intellectual

14 property. A safety taxonomy is also a mitigation here. And

15 so the -- I understand that the second bullet point, your

16 Honor, that you're focusing on here about the other safety

17 and CRS categories, I understand that that is broader. Ms.

18 Hartnett also pointed to that being broader. But with

19 respect to the post-training datasets that were used to

20 align the model to Meta's intellectual property policy in

21 bullet point one and the additional post-training datasets,

22 again sourced from the same shadow libraries alleged in the

23 complaint that were used to minimize the ability of the

24 model to output copyrighted material, those are discrete

25 datasets. Deponents have identified even places where those

17

1 datasets are located. They've named the specific shadow

2 libraries that were used to create those datasets. And even

3 Mr. Ghajar's point about these prompts, where you have

4 prompts and answers, we have testimony in this case that

5 they use the shadow libraries to generate synthetic data of

6 prompts and answers.

7   This case is about copying and use, and these are uses

8 of copyrighted material. So we really are seeking a very,

9 very narrow -- a very, very narrow set that Meta can -- if

10 Meta wants to find where these things are, it can review

11 depositions from the last two weeks of this case, and it can

12 also look at the source code repository that it produced on

13 December 2nd that pertains specifically to memorization. In

14 that repository, the source code referenced the datasets

15 that it needed to run. There is a guidepost for Meta to

16 find these things. That should not be overly burdensome.

17   THE COURT: Thank you. Would Meta like to respond

18 to that?

19   MS. HARTNETT: Yes, your Honor.

20   I believe what Counsel is referring to is Exhibit B

21 that was attached to their -- this filing, which is an

22 excerpt from the deposition of Mr. Bashlykov. And that

23 deposition testimony -- you may have well reviewed it --

24 makes clear that what that -- he is talking about is not the

25 use of data for -- and this is going back to the terms of

1 RFP 118 -- or, sorry, the -- 118 -- request 118, relating to

2 efforts or attempts or measures to prevent the models from

3 emitting or outputting copyrighted material. His -- the

4 testimony in Exhibit B is not showing a use of data for that

5 purpose, but for other training purposes. And so I think

6 the point here is that Exhibit B, I think, which is what --

7 is the only thing cited by the supplemental brief by

8 Plaintiffs for the notion that this is somehow related to a

9 copyright mitigation. That's not what the testimony in

10 Exhibit B says. And so it really is, at this point, trying

11 to cram in a request into 118 that 118 cannot support.

12   MS. POUEYMIROU: Your Honor, may I just --

13   THE COURT: I would like to hear your response.

14   MS. POUEYMIROU: Yeah.

15   THE COURT: Plaintiffs' brief, under the heading,

16 "What Plaintiffs seek," has subpoint one, "Datasets

17 identified by Mr. Bashlykov," and that's what Meta was just

18 referring to. And then subpoint two is, "Datasets

19 identified by Plaintiffs' experts via new source code." So

20 far in this hearing, we've been talking about subpoint two,

21 but now I think Meta has turned to subpoint one and has

22 argued that the testimony doesn't show that it was used for

23 the purpose referred to in RFP 118, and I do want to hear

24 Plaintiffs' response to that.

25   MS. POUEYMIROU: Well, I would direct your Honor

19

1 to page 68 and 69. I disagree with Ms. Hartnett's

2 characterization of this testimony here. Mr. Bashlykov

3 testified that -- on these pages that one of the shadow

4 libraries that has been at issue in this case since the day

5 that this case was filed was used for Meta's memorization

6 studies and mitigations. He specifically references it.

7 And that particular dataset contains every asserted work in

8 this case in its original form. And so in this particular -

9 - in this particular statement by this deponent, we already

10 know, based on what he said, that in -- 118 focuses on

11 "efforts, attempts, or measures to prevent the LLaMA models

12 from emitting or outputting copyrighted material." That is

13 memorization.

14   THE COURT: I pulled up to his testimony, and I

15 think you were directing me to pages 68 and 69. Can you

16 direct me to the lines of his testimony that you would like

17 me to focus on?

18   MS. POUEYMIROU: Yes. If you focus on from 68,

19 starting at line 17, and then you carry down to 69 and line

20 16, you'll see that he identifies a particular shadow

21 library.

22   THE COURT: Oh, I see. And that continues --

23   MS. POUEYMIROU: And, in fact, if you -- I mean,

24 frankly, your Honor, this entire -- I mean, everything that

25 follows, with respect to memorization, right? Because

1 memorization is the model's tendency to memorize its

2 training data. Regurgitation is the outputting of that

3 data, right? And so you're studying the phenomena of why

4 does the model memorize, and then you are training the model

5 to not output. To do that, there is source code, right,

6 that they've produced. It's a repository specifically on

7 memorization, and it was produced on December 2nd. And in

8 that source code, it references datasets, right, inclusive

9 of this dataset, we assume. I mean, we haven't seen it.

10 But this is not speculative. This is strong evidence, based

11 on his testimony. There was later testimony on the last day

12 of discovery by a different deponent. But in -- the general

13 course of discovery that we've gotten in this case about

14 post-training, right -- all of this falls within post-

15 training, and a model is a pre-trained and post-trained

16 model. That's the whole training process. Numerous

17 deponents have said that these datasets were used to stop

18 the model from emitting copyrighted material, and we just

19 want to see those datasets.

20   THE COURT: Thank you.

21   Let me turn to Meta. If you have a different

22 interpretation of the witnesses' testimony, why don't you

23 walk me through it, please.

24   MS. HARTNETT: We do, your Honor.

25   And, Mr. Ghajar, you should jump in, too.

21

1   There are two different concepts here. One is testing

2 the model to see if the model -- just performing tests to

3 see -- testing memorization or other concepts. A separate

4 point is whether it's training or fine-tuning the model,

5 actually affecting the model, changing the model. What this

6 request pertains to -- what they're asking for in their

7 brief is for data that was used for post-training, to train

8 and fine-tune the model. That is later discussed in Mr.

9 Bashlykov's excerpt in Exhibit B, and it's clear that that's

10 not about books. Earlier -- the point they're pointing to

11 is about whether -- is about testing, like doing testing of

12 the model to see if data that's -- you know, is anything

13 regurgitated? That is a separate issue from -- that is not

14 training the model. That's just performing tests against

15 benchmarks. And so -- the point being here that RFP 118 and

16 the way they frame the request is for information --

17 datasets used to train or fine-tune the model, and that is

18 not what they're -- if they were to ask for datasets used to

19 test the model, that would be a different request that they

20 didn't make.

21   THE COURT: Explain to me the distinction you're

22 drawing between testing on the one hand and training or

23 fine-tuning on the other?

24   MS. HARTNETT: So training and fine-tuning, my

25 understanding is that it would actually affect the model,

1 change the way the model functions. For example, feeding in

2 these question and answer prompts that are covered by the --

3 that one category, the IP category, that are not books, that

4 would actually change the way the model operates and

5 functions. That's -- that would be the post-training --

6 training, continuing to train the model. There's a separate

7 question of, how does Meta perform various tests against

8 benchmarks, just to see how the model is functioning,

9 without actually changing the model or affecting the model's

10 performance? Without -- basically, the theory here,

11 Plaintiff says that these works are being used to sort of

12 change how the model emits stuff on the back end. And our

13 point is that there's no support in the testimony that the

14 books are used for that purpose, and, at most, they're used

15 as part of a testing. And Ms. Poueymirou is correct that

16 that's what Mr. Bashlykov was talking about in those two

17 pages, and another Meta witness also said that, but they

18 already have -- those are snippets from the initial datasets

19 they already have, and that's for testing purposes, not for

20 changing the model's functioning.

21   THE COURT: Okay. Would Plaintiffs like to

22 respond?

23   MS. POUEYMIROU: Yeah. Yes, your Honor.

24   What Ms. Hartnett is referring to is validation model

25 alignment, and that is part of post-training. It's one of

23

1 the key parts of post-training. In order to have -- in

2 order for the model to do what Meta wants it to do, it

3 performs experiments, and it's -- it is a part of both the

4 pre-training phase of model development experiments on data

5 -- with data, and it is part of the post-training phase of

6 model development. Those experiments would be part of

7 Meta's efforts to stop the model from outputting copyrighted

8 material, because they are the very things that are teaching

9 Meta how bad the problem is and how to fix it. These things

10 go hand in hand, and Meta produced the memorization source

11 code. It's relevant in this case because it shows not only

12 the use of the copyrighted material, but the fact that the

13 model is likely storing a memory or copy of that copyrighted

14 material within it, and then Meta is trying to figure out a

15 way to stop it from outputting that copyrighted material.

16   To draw distinctions between testing the phenomena of

17 memorization and training the model not to regurgitate,

18 that's a distinction without a difference. It's not

19 meaningful. It all is happening within the post-training

20 phase of model development. We have 30(b)(6) testimony in

21 this case that talks about this, not only from Mr.

22 Bashlykov, but from another witness called Mr. Clark, who

23 talked about, you know, the post-training phase and how

24 books may have been used in it and how books may have been

25 used for memorization.

24

1   And so I don't think that that is a meaningful

2   distinction. We know that these training techniques that

3   Ms. Hartnett is referring to are all part of the post-

4   training stage, just as the testing, the memorization issues

5   are part of the post-training stage.

6   THE COURT: Are you saying that, first, Meta does

7   testing to see if there are problems. And then if it finds

8   that there are problems, it does fine-tuning to fix them?

9   MS. POUEYMIROU: So the testing of -- memory --

10 mitigating the model's regurgitation of copyrighted material

11 happens in many different ways, but it's guided by tests

12 that Meta is doing to see how great this phenomena is,

13 right, and then figuring out a way to stop it. And that

14 occurs in the pre-training phase. They work on mitigation

15 strategies to prevent the model from memorizing copyrighted

16 material. For example, train less on copyrighted material

17 because it will memorize more if it trains more. But in the

18 post-training stage, we know, based on the source code we've

19 gotten in this case, that Meta uses books datasets to deal

20 with both assessing the memorization issue and then figuring

21 out a solution for it. And you can't separate these things.

22 Plus, they're both part of the post-training stage of model

23 development, which we have -- frankly, your Honor, when we

24 sought in RFPs one through three -- Meta raises something

25 about RFPs one through three. When we sought the training

25

1 data for LLaMAs 1 through 3 in those RFPs, we defined

2 training data to include pre and post-training data, and

3 we've -- we have been seeking this really since the

4 beginning of this case. It's relevant. Meta has made it

5 squarely relevant because they rely on it in their fair use

6 defense, right? They say that the efforts to minimize the

7 model's ability to memorize training data verbatim, that

8 that is a critical part of how -- of their fair use of this

9 data. And so, on the one hand, they say it's extremely

10 relevant, and then on the other hand, they say, "But we

11 won't let you actually see the copies that we've made for

12 these datasets and the uses that we've used them for."

13   THE COURT: All right. Thank you.

14   Let me turn to Meta. Plaintiffs say that there's not a

15 meaningful distinction between testing and fine-tuning, that

16 those might be technically different things, but they're

17 highly related to each other. How would Meta like to

18 respond to that?

19   MS. HARTNETT: Your Honor, I think by going back

20 at first to the text of RFP 118, which is seeking documents

21 and communications, including source code, importantly not

22 training data, relating to any efforts, attempts, or

23 measures implemented by Meta to prevent the LLaMA models

24 from emitting or outputting.

25   So I think your Honor has done this previously, and we

26

1 would ask you to do it again, which is to hold them to the

2 terms of their own RFP, not like an overly hyper-technical

3 reading of it, just a plain reading of it. This does not

4 include testing data. It doesn't include datasets, and it

5 doesn't include for testing purposes. What it includes is

6 documents and communications about Meta's actually

7 implemented efforts, attempts, and measures to prevent the

8 models from emitting. And so to the extent -- the reason

9 why I was drawing that distinction is, testing is not

10 actually affecting the model's operation. It's not -- it --

11 moreover, they have testimony about it, but it's not

12 affecting the operation of the model, whereas the parts of

13 the process that they've been focused on in the LLaMA paper

14 are actual questions and answers, other things fed to the

15 model at the post-training stage to affect the model's

16 performance and function. And our point there has been that

17 those don't include books. So the thing that -- the only

18 thing they're pointing to in the record for the use of books

19 is the use of books for testing, not for affecting the

20 model's actual output or performance. That would be our

21 main point there.

22   And I would just -- on the point about the training

23 data being separately requested, your Honor has addressed a

24 similar issue in its January 2nd order. That was, I

25 believe, ECF -- I don't have the ECF right offhand, but you

27

1 had a January 2nd order where you made the point that there

2 was a RFP for training data for one through three. They

3 defined it to include pre-training. The parties negotiated

4 that, what the response was that Meta was going to provide

5 and -- post-training being largely irrelevant to what their

6 claims are in this case, which is about feeding the model in

7 the first place. And they had a deadline for moving to

8 compel on that. They didn't do that. And so we appreciate

9 that the Court is trying to be fair, but fairness is also

10 moving on the RFPs in a timely fashion that do cover

11 information and not being allowed to use RFP 118 to put in

12 anything they want now based on further assessment and

13 development on their end.

14   But bottom line is, the RFP, in our view, does not

15 cover it, and the testimony they're citing to as supposed

16 support for their theory about how the model works is

17 actually talking about testing, not about actually changing

18 the performance of the model.

19   THE COURT: Okay. Thank you. I think I

20 understand your argument.

21   MR. GHAJAR: And, your Honor --

22   THE COURT: Go ahead.

23   MR. GHAJAR: Yeah, I'm sorry. In fact, the only

24 deposition testimony you have before your Honor is what they

25 put in their motion, and it's Mr. Bashlykov testimony. And

28

1 as you resolve this issue and consider our arguments, I

2 would ask that you take a look at his testimony at page 71.

3 And I don't want to read it into the record because it's

4 designated confidential, but the conclusion that he draws in

5 response to a question from Plaintiffs' counsel is the

6 opposite of the argument you heard today. In other words,

7 the testing that he refers to is not training. And so I

8 think that's an important distinction. And so the

9 deposition testimony you have doesn't support the broad

10 reach and interpretation that Plaintiffs have argued today.

11   THE COURT: Okay. Thank you.

12   If I were to order Meta to produce the set of

13 supervised fine-tuning data that the witness identified in

14 the deposition -- let me turn to Meta -- would that be

15 burdensome? And if so, how burdensome would it be?

16   MR. GHAJAR: Ms. Hartnett -- I think there are two

17 -- there are two datasets. I think Ms. Hartnett identified

18 one. Let's see. I want to make sure that we're talking

19 about the same datasets, your Honor.

20   MS. HARTNETT: I believe -- is this the -- it's, I

21 guess, their brief at page two, and it's saying a set of

22 supervised fine-tuning data that Mr. Bashlykov testified is

23 located on -- in a specific place. I think that's what the

24 request is, your Honor?

25   THE COURT: That's the -- exactly. That's what

1 I'm asking.

2   MS. HARTNETT: Okay. We're -- let me check with

3 -- I want to make sure I'm accurate about -- if you could

4 just give us a moment. We're consulting with our team to

5 make sure we make an accurate representation about burden

6 for that.

7   THE COURT: That's fine. And then, in the

8 meantime, let me turn to Plaintiffs.

9   Under bullet point two, when you say, "Datasets

10 identified by Plaintiffs' experts via new source code" --

11 we've talked about your first bullet point and the second

12 bullet point. The third bullet point is, "Any additional

13 post-training datasets sourced from shadow datasets and used

14 by Meta to fine-tune its LLaMA models to minimize their

15 ability to memorize or output training data verbatim." When

16 I read that, I don't have a sense of what exactly that

17 bullet point is asking for. And, in particular, I don't see

18 how it's different from your first bullet point under item

19 number two, but perhaps you could let me know if you think

20 -- it sounded more like a catchall and like you weren't

21 really sure if there's anything else there, but I wanted to

22 give you a chance to respond.

23   MS. POUEYMIROU: So, thank you, your Honor. There

24 is actually a distinction between the first and the third

25 bullet points here. So, for the first bullet point, it is

30

1 -- you know, again, we don't have the evidence, but based on

2 what we know, that particular dataset, the datasets that

3 could be used to fine-tune so that the model is aligned with

4 that intellectual property category, we believe they're

5 sourced from the copyrighted books at issue in this case,

6 but it would be a different kind of dataset, potentially,

7 than the datasets that would be used to minimize

8 memorization. One would be focused on how to respond to end

9 users. The other one would also be focused on responses to

10 end users, but also on how to deal with actually lessening

11 memorization of the model. And that third bullet point is

12 actually very -- it was written -- it is aligned with Meta's

13 statement on its fair use interrogatory response, that its

14 efforts to minimize the model's ability to memorize and/or

15 output training data verbatim are -- is evidence of fair use

16 here, probative of the question of fair use.

17   There is source code that Meta has produced in this

18 case that focuses solely on the issue of memorization. And

19 in that source code, datasets are referenced. Because Meta

20 was concerned with the model memorizing personal identifying

21 information as well as copyrighted information, it would

22 have used different datasets to effectuate those types of

23 mitigations. And so what this is specifically asking for is

24 for the memorization source code that's been produced in

25 this case, that focuses exclusively on the model memorizing

31

1 copyrighted books. We would want to see that small subset

2 of post-training datasets that Meta used for that source

3 code it has produced.

4   THE COURT: The first bullet point seems like it

5 describes something with at least some precision, and the

6 third bullet point doesn't seem to have that same type of

7 precision. Can you help me out more with the --

8   MS. POUEYMIROU: So the precision -- the -- I

9 think the way to understand how this is actually a very

10 narrow request is that Meta wants its LLaMA models to be

11 able to generalize knowledge, right? But, you know, large

12 language models also memorize their training data and leak

13 it. It's a problem of overfitting. And I -- in order to

14 prevent the model from just leaking training data in

15 response to end users, there is source code that pertains to

16 that particular phenomena with respect to copyrighted

17 material. So how do we prevent the model from memorizing

18 and regurgitating the copyrighted books it's trained on?

19 There's source code. There's an entire source code

20 repository that's focused exclusively on memorization. And

21 in that repository, there is stuff that deals with privacy,

22 right? If you train on personal identifying information,

23 you don't want the model to leak it. So Meta has got

24 datasets that are going to deal with that.

25   But it also has datasets and source code that deals

1 with specifically IP concerns. And this is not speculation.

2 This is in the discovery we've gotten to date where Meta

3 talks about IP leakage, about the leakage of copyrighted

4 books from the -- on -- from the training data on which the

5 LLaMA models have trained. So what we are seeking here

6 specifically is the training datasets that Meta used in its

7 efforts to minimize the model's ability to memorize or

8 output that copyrighted training data. It's actually

9 specifically the training data that Meta is citing in its

10 own interrogatory response that helped it effectuate the

11 thing that it claims that it's been able to effectuate.

12   THE COURT: So if I understand the third bullet

13 point, training data could also include personally

14 identifiable information. And that's not really at issue in

15 this case. That's a separate concern from copyrighted

16 material. But if the third bullet point is limited to

17 datasets that prevent the memorization or output of

18 copyrighted material, why isn't that the same as the first

19 bullet point?

20   MS. POUEYMIROU: Because the first bullet point is

21 also potentially a fine-tuning mechanism. So a training

22 dataset of answers and responses -- sorry, of prompts and

23 responses that are based on copyrighted books are pulled out

24 from copyrighted books to teach the model how to respond to

25 end users. And that is about how the model outputs

33

1 responses when asked to provide -- if someone asks the model

2 to give it a copyrighted book, like Harry Potter, right, the

3 model will have a script of how to respond to it. That is

4 potentially the type of dataset that is being used to align

5 the models with this intellectual property category. Based

6 on testimony we've gotten in this case, they use copyrighted

7 books for those prompts and responses datasets. But that's

8 one specific kind of fine-tuning, your Honor. And the

9 datasets that are referenced in this third bullet point,

10 it's more expansive. It's still narrow, because it's just

11 -- we're just interested in the memorization datasets that

12 are copyrighted, you know, that comprise copyrighted

13 material, but it is potentially a different kind of training

14 or fine-tuning. They're like -- they're part of the same

15 species. They're just different types of approaches here,

16 but all use what we believe to be dataset source from the

17 shadow libraries at issue in this case that contain our

18 Plaintiffs' works.

19   THE COURT: Let me turn to Meta. On this third

20 bullet point, under item number two, if we narrowed it from

21 training data to copyrighted material, what are your

22 thoughts on whether those post-training datasets are as

23 Plaintiffs describe?

24   MR. GHAJAR: I'll start, and Ms. Hartnett should

25 chime in.

1   I don't -- we don't believe they are as described, and

2 we wouldn't know where to begin to look for them. They

3 haven't been specifically identified, your Honor. We don't

4 know what they're referring to. I don't. I did a lot of

5 diligence in preparation for this hearing. It's one thing

6 when your Honor refers to bullet point one under Bashlykov,

7 which, for the record, does not have to do with IP

8 mitigation, but at least it's identified. And in bullet

9 point one, under subsection two, at least there is a

10 reference to CRS 17 intellectual property. We know where to

11 look for it, in spite of our position that it's not relevant

12 or responsive. I don't -- and we don't know where to even

13 begin to indulge the speculation by Counsel as to certain

14 datasets that supposedly were used and are supposedly

15 referred to in the source code. It's -- it -- that is --

16 does not comport with our understanding, your Honor.

17   THE COURT: Okay. Thank you.

18   And then let's turn back to item number one. Meta was

19 conferring with team members to answer the question about

20 burden. Do you have an answer on that at this point?

21   MR. GHAJAR: As to the dataset under section one,

22 Mr. Bashlykov -- again, setting aside arguments as to

23 relevance and scope, that could be produced in a matter of a

24 couple of weeks.

25   Ms. Hartnett, jump in if I misspeak. If --

1   MS. HARTNETT: I think that one appeared to be

2 several weeks, due to the size of it. It's -- I'm not sure.

3 We don't have the number of gigabytes, but that one is a

4 larger set, so that would take -- I think several weeks is

5 our best -- what we're able to say right now.

6   THE COURT: Okay. Thank you for the information.

7   And then let me turn back to Plaintiffs. Under

8 subheading number two, the third bullet point that we've

9 just been talking about, Meta says it doesn't know what

10 you're referring to and wouldn't know where to begin to

11 look. I would like to hear your response to that.

12   MS. POUEYMIROU: I would direct Meta to its source

13 code repository MME. It produced that in this case on

14 December 2nd. And that source code repository is a

15 compendium of all of the source code since -- we assume that

16 Meta has put together and that exists to effectuate Meta's

17 interest in preventing memorization and preventing the

18 regurgitation of training data.

19   Meta doesn't want its models regurgitating Wikipedia

20 posts, it doesn't want its models regurgitating PII, and it

21 doesn't want its models regurgitating Harry Potter. In

22 order to stop its models from doing those three different

23 types of things, it uses datasets to achieve it. And the

24 source code references those datasets. We -- our experts

25 have been asking for these datasets for months. The source

1 code is expressly referencing them as well, your Honor, and

2 they are -- what Mr. Ghajar could do is have -- could focus

3 on this specific repository and could focus specifically on

4 just the datasets that Meta used for its copyright

5 memorization mitigations. It has produced discovery in this

6 case where it divided its memorization mitigations into

7 three prongs. One was just general leakage of training

8 data, because it wants its model to be able to generalize

9 knowledge, one was the concern about privacy, and one was

10 the concern about copyright. It referred to copyright books

11 datasets there. Those are the datasets that we assume Meta

12 is using, alongside the source code, in its memorization

13 repository. Very small, discrete datasets. I don't know

14 the size of the datasets, but it is a small number of

15 datasets that we think they are using. And we also know

16 that its copyrighted books datasets are sourced from the

17 shadow libraries at issue in this case.

18   THE COURT: What's your reason for thinking that

19 the post-training datasets contain the copyrighted

20 materials? I understand your argument that the shadow

21 datasets contain the copyrighted materials, but you seem to

22 also be making the argument that the post-training datasets

23 that are sourced from the shadow datasets, that those post-

24 training datasets, as well, contain the copyrighted

25 material. What's your reason for thinking that?

1   MS. POUEYMIROU: Is the question whether they

2   contain Plaintiffs' asserted copyrighted materials or just

3   copyrighted books as specific copyrighted materials?

4   THE COURT: Oh, that's a fair point. Let's take

5   the latter one, whether they have any copyrighted books.

6   MS. POUEYMIROU: So the memorization mitigation

7   techniques that Meta is doing as part of these efforts, the

8   copyrighted mitigation techniques are specifically

9   copyrighted books. It's not a -- it's not copyright writ

10 large. It's a copyright book memorization issue. They are

11 unsurprisingly concerned about the model memorizing

12 copyrighted material and regurgitating that material. So

13 it's a discrete project at Meta. In the discovery that I'm

14 thinking of, in documents that have been produced in this

15 case, when discussing the copyrighted books mitigation

16 strategies, the shadow libraries at issue in this case are

17 referenced there, said, pulled books from these places to

18 effectuate this particular mitigation technique. And so in

19 the source code repository, the MME source code repository

20 that was produced on December 2nd, the copyrighted books

21 mitigation techniques should be their own thing inside of

22 that repository, because our source code expert has

23 confirmed that they are. And the datasets that were used

24 with that source code would be what we're seeking.

25   THE COURT: And if I were to change the question

38

1 to ask whether these post-training datasets contain the

2 Plaintiffs' copyrighted books, would your answer be that you

3 don't know until you see the datasets?

4   MS. POUEYMIROU: My answer would be we think it is

5 highly plausible that they do, because we have identified in

6 the training data Meta has produced in this case from both

7 of these shadow libraries our Plaintiffs' works in these

8 datasets. And so to the extent we also know that it was

9 those same shadow libraries that Meta sourced from for its

10 copyrighted book mitigation efforts, we feel that it is a

11 highly plausible inference that you could draw that

12 Plaintiffs' asserted works were among those later works in

13 those datasets.

14   THE COURT: In a prior order, I said that the term

15 "shadow datasets" was a matter of opinion and that Meta

16 should not be forced to guess what you think is a shadow

17 dataset. How do we deal with that with your third bullet

18 point?

19   MS. POUEYMIROU: So Meta actually -- if Meta wants

20 to run the term shadow library in the discovery it has

21 produced to date -- it produced, in its latest round of

22 discovery, a document that cataloged all of the libraries it

23 considered to be shadow libraries. It can look to its own

24 discovery for that. With respect to -- I think Meta knows

25 which libraries we're talking about, but the discovery

39

1 that's been produced in this case is very clear about which

2 shadow libraries are at issue here.

3   THE COURT: Is there a way of rephrasing your

4 third bullet point that it doesn't use the word "shadow

5 datasets," given my prior order on that issue?

6   MS. POUEYMIROU: I mean, we could specify, which

7 would align with RFPs we issued right in the beginning of

8 this case -- we named all the shadow libraries. And so we

9 could specify -- I think those are RFPs six through 10. I'm

10 not sure. But we could expressly name which ones we're

11 discussing here.

12   THE COURT: I see.

13   MS. HARTNETT: Your Honor, we -- I think you'll

14 give us a chance to respond, but we do have a response on

15 this point.

16   THE COURT: Let's have it. And why don't you

17 respond to all of the points that the Plaintiffs have made?

18   MS. HARTNETT: I will do my best.

19   I think the most important point is -- and we could

20 talk about the time it would take to produce what's under

21 the Bashlykov testimony and maybe even the specific IP

22 mitigation category, but going to bullet point three, and

23 what Counsel is now proposing is a massive expansion. We

24 don't understand exactly what she's referring to or talking

25 about. None of this has been briefed. As she herself has

1 said, they have the source code since December 2nd, and the

2 briefing -- your Honor has given them a chance to be more

3 specific in the briefing about what they're asking for.

4 This is the first time we're hearing about references in the

5 source code to potentially cross-referencing datasets. This

6 is a massive source code repository that we would have to, I

7 guess, go through. It would -- again, I want to make sure

8 I'm making accurate representations, but it would be months

9 to go through and see if there's -- these are not just,

10 like, hyperlinks in the source code to datasets -- to figure

11 out, is there anything in here that might relate to what

12 they're talking about?

13   So, again, I just really appreciate that you're trying

14 to hear it out and be fair. But on the other hand, we're so

15 far afield from not only 118, but then -- that was about the

16 LLaMA papers, and then it was about -- supposed to be about

17 narrowing and asking for something specific in the LLaMA

18 paper. And now I understand the request to be actually to

19 go through the entire source code repository and see if

20 there are links to datasets. And this was not briefed.

21 It's not something we've had a chance to fully tell the

22 Court the burden, but I can tell you it's an alarming level

23 of burden, what they're suggesting.

24   It also seems to -- it just is not connected to, at the

25 end of the day, what they've asked for in 118, and it's

41

1 something they should have brought to the Court's attention

2 sooner if it was so important. This is really unfair to us

3 to try to defend against that request at this hearing.

4   For the other ones, we will -- you know, we -- I think

5 we've made our points about relevance and burden, and, you

6 know, the most discrete request that they have would be,

7 although irrelevant because it doesn't involve books, the CR

8 -- I don't want to be -- I want to be careful about how I

9 describe things with the protective order -- would be the

10 first bullet under two. I think that's the most discrete

11 and something that would be doable on the timeframe of the

12 case. The Bashlykov-testified information, again, not

13 relevant, but would take longer, probably several weeks, to

14 pull that. And then for the second and third bullet points

15 under two, those are just such broad -- I mean, it just --

16 they're very broad and indefinite, and it apparently would

17 require us to go through this entire source code to come up

18 with what may be cross-references to datasets that are

19 speculative.

20   MS. POUEYMIROU: Your Honor, just one point on

21 this enormous source code. I did want to say that Meta did

22 dump an enormous amount of source code. It dumped all the

23 source code for Facebook and WhatsApp on us, 3.8 terabytes,

24 on December 4th, I think, but it also provided a discrete

25 repository just on memorization. And that -- that's not an

42

1 enormous repository. That is a repository focused just on

2 Meta's efforts to minimize the model's ability to memorize

3 -- outputting its training data. And that repository has

4 privacy -- a privacy area, and it has a copyrighted books

5 area as well, from what our source code experts have told

6 us. And he has also told us that in that repository, with

7 respect to that specific source code, datasets are

8 referenced.

9   So I don't think that it is actually a needle in a

10 haystack sort of endeavor here. I think it's actually much

11 more manageable. And we've been -- you know, we have been

12 trying to seek this data for a long time, so I -- we would

13 be more than happy to -- you know, if it would help, we

14 could have our source code expert specifically identify

15 parts of the source code. We could go into Cooley's office

16 and do that, if that was helpful. But I think it is

17 actually something that Meta will be able -- I mean, just

18 speaking to the engineers that we deposed in this case,

19 specifically, actually, Mr. Seobu (phonetic), the final

20 deponent you had, who works on memorization for Meta, who we

21 deposed on December 14th, I think he would probably be able

22 to identify where these datasets are stored.

23   THE COURT: All right. Well, thank you, Counsel.

24   MR. PRITT: Just real quick, your Honor. Just in

25 case it's helpful to Meta, many of these specific

43

1 repositories in the source code are also mentioned in the

2 declaration of Doctor Jonathan Krein, who's our source code

3 expert from November 7. So they can also --

4   MS. HARTNETT: Your Honor, may -- if I just may on

5 the document on point, just for the public record, since

6 this is a public hearing. There have been a lot of

7 representations that are just not accurate or overblown. On

8 that, in particular, that source code was produced because

9 -- it was in response to their request for -- we gave them

10 the entire source code, so -- because it's not able to be

11 parsed out for something that -- specific that they wanted.

12 So that was a good thing, not a bad thing, not a dump. But

13 even aside from that, I just think these arguments that

14 we're hearing now about specifics this and that are things

15 that should have been made to us long ago and not as a last

16 ditch effort at this hearing to try to get information that

17 was never specifically requested, was never subject to a

18 meet and confer, and now is -- that was never part of the

19 supplemental briefing that you allowed the party to be able

20 to specify what they want, and this is the first time we're

21 hearing a lot of this.

22   THE COURT: All right.

23   Counsel, I thank you for your arguments. I feel that

24 I've heard sufficient arguments from both sides that I'm

25 able to rule on the discovery dispute that's in front of me.

44

1 And so I will get out a written order on -- resolving the

2 joint discovery letter brief. And with that, the matter is

3 submitted.

4   MS. POUEYMIROU: Thank you, your Honor.

5   MR. PRITT: Thank you. Have a good day.

6   MR. GHAJAR: Thank you, your Honor.

7   THE CLERK: Thank you, everyone.

8   MR. PRITT: Thank you, everyone.
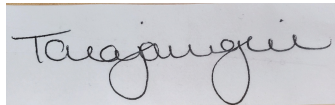
9   THE CLERK: We're off the record in this matter,

10 and court is in recess.

11   (Proceedings recessed at 10:36 a.m.)

12

13

14

15

16

17

18

19

20

21

22

23

24

25

45

1   CERTIFICATE OF TRANSCRIBER

2

3   I certify that the foregoing is a true and correct

4 transcript, to the best of my ability, of the above pages of

5 the official electronic sound recording provided to me by

6 the U.S. District Court, Northern District of California, of

7 the proceedings taken on the date and time previously stated

8 in the above matter.

9   I further certify that I am neither counsel for,

10 related to, nor employed by any of the parties to the action

11 in which this hearing was taken; and, further, that I am not

12 financially nor otherwise interested in the outcome of the

13 action.

14

15

16   Echo Reporting, Inc., Transcriber

17   Sunday, January 12, 2025

18

19

20

21

22

23

24

25